EXPLORATORY ANALYSIS OF ZOOPLANKTON SPECTRA USING MATRIX DECOMPOSITION TECHNIQUES

Nikolai Sushkov[1]*, Timur Labutin[1], Nikolai Lobus[2], Gábor Galbács[3]

[1]Department of Chemistry, Lomonosov Moscow State University,
119234 Moscow, Leninskie gory 1 str. 3, Russia
[2]Shirshov Institute of Oceanology of the Russian Academy of Sciences,
119997 Moscow, Nakhimovskii pr. 36, Russia
[3]Department of Inorganic and Analytical Chemistry, University of Szeged,
6720 Szeged, Dóm square 7, Hungary
**e-mail: nikolaisushkov@laser.chem.msu.ru*

1. INTRODUCTION

Marine zooplankton communities of high latitudes are often dominated by copepod crustaceans belonging to the order *Calanus* (Fig. 1). They are a link between primary producers (phytoplankton) and higher levels of the food chain, thus making a significant contribution to energy flows in the marine ecosystem. A characteristic feature of calanoid copepods is the so-called winter diapause, which is accompanied by hibernation at a great depth (more than 500 m). In spring, the crustaceans migrate back upwards. The mechanism of buoyancy regulation is not completely clear. It may be based on a change in the density of lipids contained in a special fat bag, depending on the water pressure during vertical migration. Body density can also be regulated through the exchange of heavy ions for lighter ones (e.g. Na^+ for NH_4^+). In addition to ammonium ions, lithium ions can also reduce the density. The Li content in seawater is 28 mmol/l (about 200 mg/l), while in calanoids it can be 2-3 orders of magnitude higher. Accumulation of lithium through the food chain and passive uptake from the environment are considered unlikely. Thus, it should be assumed that there is a physiological mechanism for active accumulation and that lithium has some biological function which is yet to be clarified [Lobus 2016, Freese 2015, Lobus 2018].

Figure 1. *Calanus glacialis* (the scale bar is 2 mm)
[http://www.arcodiv.org/watercolumn/copepod/Calanus_glacialis.html, retrieved Nov. 09, 2020].

The main components of copepod tissues in terms of molecular composition are proteins, fats, and chitin [Freese 2016]. There is also intense coloration due to the presence of carotenoid pigments, whose signals are clearly visible in absorption and/or Raman spectra (e.g. [Zagalsky 1985]). The goal of the present study was to find correlations between atomic and molecular composition of copepod tissues to provide insight into reasons behind lithium accumulation. To this end, analyses by LIBS and Raman spectroscopy were performed.

2. EXPERIMENTAL

Samples. We analyzed 29 zooplankton samples, 14 of which were calanoid copepods. The animals were caught during expeditions in the Arctic seas and in the Black Sea in 2014–2017 (typically from August to October). They were washed with deionized water, dried at 50° C for 12 hours and pelletized under a moderate pressure (d = 8 mm, 20–1000 bar, typically 30 bar).

Equipment. LIBS was performed using the Applied Spectra J200 Tandem LA-LIBS instrument (excitation laser: 266 nm Nd:YAG, 20 mJ/pulse, 10 Hz, spot diameter: 200 μm; acquisition delay: 500 ns). Since the instrument was equipped with a CCD detector, we obtained time-integrated spectra (integration time ~ 3 ms). The spectra covered the entire optical range (186–1049 nm). Resolving power depended on the wavelength range and was typically around 3000.

Raman scattering spectra were recorded using the Thermo Scientific DXR Raman Microscope (780 nm laser, energy 1-14 mW, spot size 1–2 μm) in the range of 45–3500 cm^{-1}. We applied different energies depending on the fluorescent background intensity and resistance of samples to burning. Aperture was selected individually for each sample to attain sufficient signal-to-noise ratio and avoid CCD overflow at the same time. The following apertures were used: 25 and 50 μm slits, and 25 and 50 μm pinholes. Acquisition time ranged from 30 s to 5 min. Spectra were instantaneously fluorescence corrected in the instrument software using a 6th order polynomial function.

Experimental procedures. Due to inhomogeneity of the samples, we distinguished dark, light, and medium coloured spots on the pellet surfaces. For each of the colours, LIBS and Raman spectroscopy were carried out in at least 3 spots of that colour, i.e., a total of at least 18 spots were examined on each pellet by these techniques. In the case of LIBS, typically 10 laser shots per spot were delivered.

2.1. DATA PROCESSING METHODS

a. Data pretreatment. The obtained arrays of emission spectra were filtered for outliers. To this end, several prominent peaks, including analytical lines, were selected. The list typically included peaks at 610.34 (Li I and Ca I), 247.84 (C I), 473.60 (C_2 band head), 568.30 (Na I) and 396.85 (Ca II) nm. Their background corrected heights were checked for outliers using the Grubbs' criterion, which was applied repetitively until there were

no outliers at any peak. This procedure was supervised in a home-made graphical interface, so that results could be corrected if necessary. The remaining spectra were averaged for each of the zooplankton samples, and an array of averaged spectra (29 samples × 12 275 wavelengths, 186–1049 nm) was composed. We performed no background correction for emission spectra.

Raman spectra were first manually background corrected since the automatic correction was not always satisfactory. Then we divided each spectrum by its mean to reduce the effect of different experimental conditions. After that, the spectra were averaged. Dimensions of the resulting matrix were 29 samples × 2801 wavenumbers (450–3150 cm^{-1}).

Bulk composition data (elements from Li to U except non-metals), obtained by ICP-AES and ICP-MS after digestion, were also available.

Data processing was carried out in Origin 8.5, GNU Octave, Microsoft Excel, Wolfram Mathematica 8, and Matlab 2020 software. In order to perform an exploratory analysis of the data, we used various matrix decomposition techniques described below.

b. Principal component analysis (PCA). Decompositions by PCA were a starting point for all further considerations. To determine the optimum number of principal components, we used conventional scree plots of eigenvalues. The same number of components was implied in NMF decompositions (see below).

c. Non-negative matrix factorization (NMF). A detailed discussion of this technique, which may be considered as one of the methods for blind source separation, can be found in [Cichocki 2009]. The approach is a matrix factorization constrained by assumption that resulting components should be non-negative. This helps to obtain physically interpretable components. The problem may be formulated as follows: given a matrix X with m descriptors (rows) and n samples (columns), factorize it into two terms W ($m \times n$) and H ($n \times p$), with $p < \min(m, n)$:

$$X = WH + E$$

Here, the matrix E represents approximation error. Columns of W are called basis vectors, and rows of H are decomposition coefficients. Contrary to other decomposition techniques like PCA, the additive nature of NMF does not allow factors to vanish due to their equal values and opposite signs. The technique has become popular in several areas, such as facial recognition and processing of electroencephalographic data. The drawback of NMF is that the solution is not unique since optimization only leads to a local minimum of cost function. Additionally, the results depend on the initial approximation of W and H. Therefore, it is advisable to perform multiple (10–1000) algorithm runs to obtain a reliable factorization [Buciu 2008, Brunet 2004].

d. ComDim—PCA and ComDim—ICA. The common component and specific weights analysis, also known briefly as ComDim, was first introduced in 1995 [Qannari 2004]. It deals with several blocks of data (e.g., LIBS and Raman data) and seeks to factorize them independently, without concatenation. To this end, the ComDim defines underlying common dimensions relevant to each block and assesses how much each dimension is

relevant to each data block. The measure of the "relevance" is called specific weight, or salience.

If each of N samples is represented by a line in a block of measurements X_k, it is possible to define the W_k matrix as a scalar product $X_k X_k^T$. This matrix is modelled as

$$W_k = Q\Lambda^{(k)}Q^T + E_k$$

Here Q is an orthogonal matrix of dimensions (N, N), which contains the common components in its columns. The matrix $\Lambda^{(k)}$ is a diagonal (N, N) matrix of specific weights; it can be different depending on the block X_k, while Q is the same for all the blocks X_k. Sum of saliences in $\Lambda^{(k)}$ describes to what extent a particular X_k is contributing to Q.

ComDim components are calculated one by one. These steps may be implemented to involve PCA or ICA (independent component analysis). Contrary to PCA, ICA does not assume orthogonality of components, which brings it closer to reality, but at a cost of solution being non-unique. Like NMF, it is a blind source separation technique and seeks to find components as independent as possible by maximizing a certain function which defines the independence [Hyvärinen 1999].

3. RESULTS AND DISCUSSION

3.1. The whole dataset

We have first investigated the whole dataset which contained data about all the 29 samples of crustaceans. We have concatenated LIBS spectra to obtain matrices containing data from dark, medium-coloured, and light spots (29 samples × 36825 predictors). In this fashion, all the available information is used. PCA decomposition of this matrix yielded 5 components (explained variance 99%). One of the obtained loadings contained emission signals of Li, Na, and Mg. Calanoid crustaceans formed a group on the score graphs, which was however not well separated from the other samples. To improve this result, we cut out all resonance lines from the data matrix (like C I 247, Na I 589, K I 769 nm, etc.) to avoid nonlinear signals. Spectral range with wavelengths longer than 947.8 nm was also discarded. There was clearer clustering in score graphs obtained for this shortened matrix than in the previous case.

Same operations were tried with concatenated Raman spectra (8403 predictors). PCA decomposition yielded 5 PCs, but loadings were not easy to interpret, and no clear clustering was observed on score graphs. Discarding wavenumbers greater than 1921 cm^{-1} helped to avoid strong and non-characteristic hydrocarbon chain signals around 3000 cm^{-1} and a large uninformative area between 1921 and 2700 cm^{-1}. Score graphs obtained from the shortened matrix allowed to distinguish crustaceans (incl. *Calanoida*) from other animals (sea snails and arrow worms). However, it was

impossible to distinguish calanoids from other crustaceans according to their Raman spectra (in contrast to the LIBS results).

LIBS and Raman spectra can also be concatenated together to give a hybrid dataset. For these data, both PCA and NMF gave 2 "Li-containing" loadings. In the Raman domain, one of them contained signals of amino acids like tryptophan (758 cm^{-1}), and the other one showed carotenoid peaks. Both methods provided reasonably good clustering allowing to distinguish calanoids from other crustaceans and snails, but not from arrow worms (**Figure 2.**).

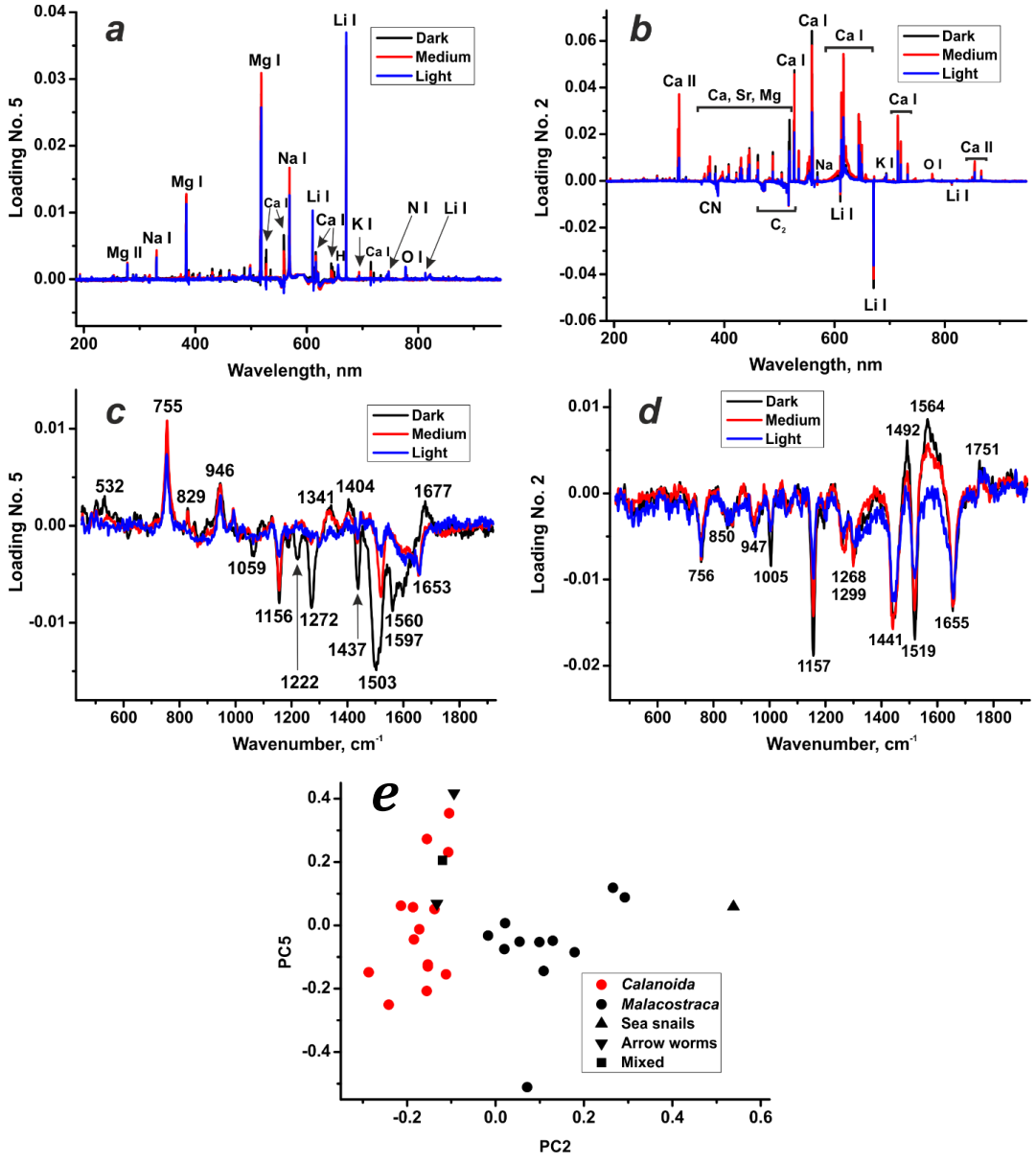

Figure 2. Results of the PCA decomposition of the combined LIBS+Raman dataset (loadings Nos. 2 and 5); a, b: LIBS domain; c, d: Raman domain; e: score graph (PC5 vs. PC2). Calanoid copepods are shown in red.

We also compared performance of ComDim—PCA and ComDim—ICA regarding our data. In both cases there were 6 separate blocks of data, corresponding to 3 spot colours in LIBS and Raman spectra. Spectra were shortened as described above.

Although both kinds of ComDim revealed correlation between Li signals, carotenoid and amino acid bands, loadings were easier to interpret in ComDim—ICA. It should be noted that the correlation with Trp band is stronger in dark and light spots compared to medium-coloured ones. Classification was better in ComDim—PCA.

3.2. *Calanoida* samples only

This section deals with the data subset related to 14 samples belonging to the *Calanoida* order. These spectra were pre-treated as described in the previous section (resonance lines discarded, etc.). PCA and NMF decomposition of the LIBS spectra readily gave a component dominated by Li signals, and on the score graphs, samples showing anomalously high Li content tended to be located separately from other samples. Raman spectra yielded separate loadings corresponding to carotenoids and to amino acids. This trend was more distinct for the PCA than for NMF. The Raman score graphs are less informative than the LIBS-related graphs; classification is apparently driven mainly by the intensity of carotenoid bands.

As for combined LIBS+Raman data, both PCA and NMF mainly underline the correlation between Li signals and carotenoid bands, although bands of amino acid are also present in the relevant loadings. There is reasonably distinct clustering of samples on the score graphs according to their Li content.

ComDim—PCA (**Figure 3.**) again placed Li together with carotenoid pigments and amino acids. It is interesting that ComDim—ICA suggested 2 carotenoid-containing loadings with roughly equal proportions of explained variance. Lithium signals dominated in the loading which contained more tryptophan signal. This may suggest that Li accumulation is more importantly related to amino acids than to pigments.

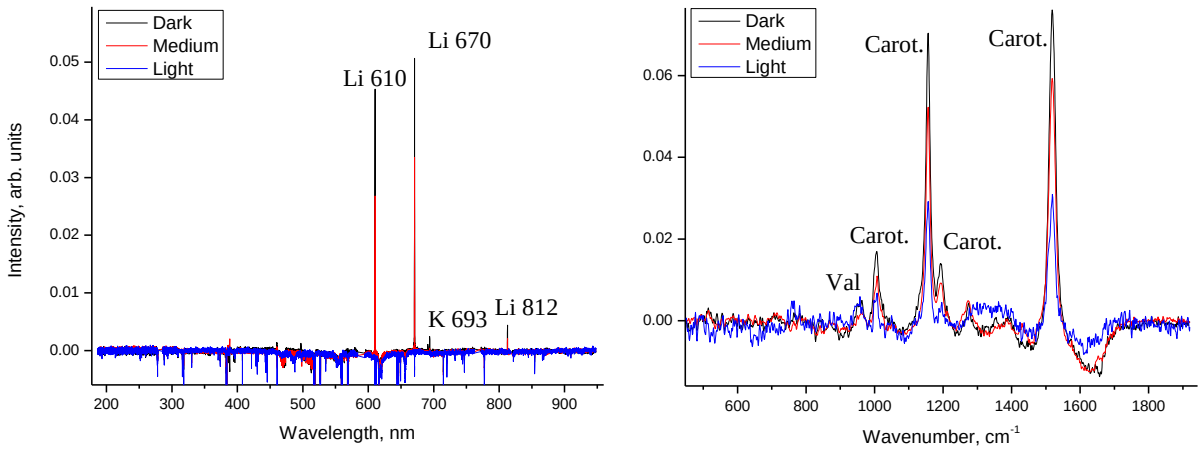

Figure 3. The Li-related common component in ComDim. Left: LIBS data. Right: Raman data. "Carot." = carotenoid compounds, "Val" = valine.

Indeed, there is a distinct pair correlation between Li bulk content (as determined by ICP-MS) and the intensity of Trp peak in Raman spectra (**Figure 4.**). There is no comparable correlation between Li and carotenoid bands.

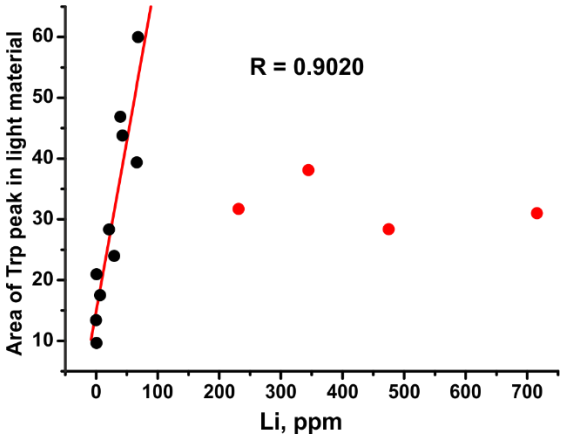

Figure 4. Linear correlation between Li content and Trp band in Raman spectra of calanoid copepods. Samples marked in red are considered anomalous and not included into the regression.

4. CONCLUSIONS

Matrix decomposition techniques used in this study for exploratory analysis of LIBS and Raman spectra revealed correlation between Li content, amino acids, and carotenoid pigments in marine zooplankton. The most interesting correlation is that with tryptophan. Results obtained by PCA, NMF, ComDim—PCA, and ComDim—ICA are generally the same, although PCA-based approaches are often more precise in classification and yield higher explained variance. Discarding resonance emission lines and uninformative regions of LIBS and Raman spectra helped to improve results of matrix decomposition. The most productive way of analysing LIBS and Raman spectra is data fusion, either by simple (low-level) concatenation or by more sophisticated techniques such as ComDim.

5. ACKNOWLEDGEMENTS

ICP-AES and ICP-MS elemental analyses of zooplankton samples were funded by the Russian Science Foundation (research project No. 18-77-00064). The authors are grateful to Á. Bélteki, Dr. A. Kéri, P. Janovszky, D. Palásti, Dr. K. Fintor (University of Szeged) and Dr. R. Rajkó (Universtity of Pécs, Hungary) for their valuable assistance. The financial support received from the National Research, Development and Innovation Office of Hungary through project No. K_129063 is kindly acknowledged. The MATLAB code for performing ComDim was kindly provided by Prof. D. N. Rutledge (AgroParisTech, France).

6. REFERENCES

[Brunet 2004] J.-P. Brunet, P. Tamayo, T. R. Golub, J. P. Mesirov, *Proc. Natl. Acad. Sci.*, 101 (2004) 4164.

[Buciu 2008] I. Buciu, *Int. J. Comput. Commun. Control*, 3 suppl. issue (2008) 67.

[Cichocki 2009] A. Cihocki, R. Zdunek, A. H. Phan, S. Amari: Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation, *Wiley Publishing*, 2009.

[Freese 2015] D. Freese, B. Niehoff, J. E. Soreide, F. J. Sartoris, *Limnol. Oceanogr.*, 60 (2015) 2121.

[Hyvärinen 1999] A. Hyvärinen, *Neural Comput. Surv.*, 2 (1999) 94.

[Lobus 2016] N. V. Lobus, *Oceanology*, 56 (2016) 809.

[Lobus 2018] N. V. Lobus, A. V. Drits, M. V. Flint, *Oceanology*, 58 (2018) 405.

[Qannari 2004] E.M. Qannari, I. Wakeling, H. J. H. MacFie, *Food Qual. Prefer.*, 6 (1995) 309.

[Zagalsky 1985] P. Zagalsky: Invertebrate carotenoproteins (Chapter 9) in: Methods in Enzymology, Steroids and Isoprenoids Part B (ed. by J. H. Law, H. C. Rilling), *Elsevier*, 1985.